PHIL 7001: Final Exam

Professor Boris Babic

Instructions:

This is the final examination for PHIL 7001. The assignment will be graded out of a total of 100 points. It is closed book, but you can have one double spaced cheat sheet, on which you can write anything you like.

You have 180 minutes to complete this test. If you finish early, you may submit your answers to the test invigilators and leave early. However, in order to avoid too much commotion, during the last 30 minutes please do not leave. That is, if you finish with less than 30 minutes remaining, please remain at your desk quietly until the exam is over.

Please make sure to write as clearly and legibly as possible. This will make it much easier for grading purposes.

You can write your answer in the exam booklet itself. Directly below each question. If you need more space, you can use the back side of each paper of the exam booklet.

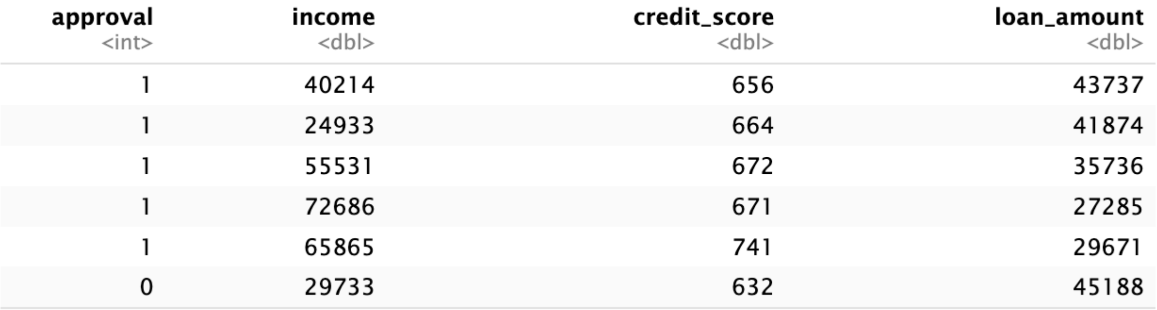
Calculators are allowed, and you can use any calculator you wish. Smart devices of any kind are not allowed.

Please write your name and student number clearly on the first page of your submission.

There should be no trick questions on the exam. If anything seems unclear or ambiguous, then you can use your judgment to resolve the ambiguity, and note that you have done so in your answer. It is always best to state your assumptions as clearly as possible.

Problem 1: Logistic Regression (35 Points)

Suppose that you are working with a dataset called 'loan_data.csv', which contains information about loan applications. This dataset includes information such as an applicant's annual income (in dollars), credit score, and loan amount (in dollars). Your goal is to build a logistic regression classifier that predicts whether a loan application will be approved or denied. This is a binary classification task where '1' represents approval, and '0' represents denial. For the purpose of this analysis, we will assume a statistical significance level of 5%. Below you can see the first several rows of this dataset.



approval <int>	income <dbl>	credit_score <dbl>	loan_amount <dbl>
1	40214	656	43737
1	24933	664	41874
1	55531	672	35736
1	72686	671	27285
1	65865	741	29671
0	29733	632	45188

1A. Before fitting the classifier, your first task is to split the loan data into a training set and a testing set. Suppose we use 20% of the data as our training set, and 80% of the data as our test set. Is this a good idea? Explain in 1-2 sentences why or why not. (3 points)

It's not a good idea because using only 20% of the data for training means that the model is trained on a very small subset, which might not be representative of the entire dataset. This could lead to poor model performance due to underfitting. Typically, a larger portion, like 70-80%, is used for training to ensure the model learns the underlying patterns well.

1B. Suppose we've split the loan data into a training set 'loan_train', and a testing set 'loan_test'. You then need to fit a logistic regression model to predict loan approval, where income and credit score are the predictors. Please complete the blanks. (4 points)

```
logistic_model <- glm(approval ~ income + credit_score,
                      data = loan_train,
                      family = ‘binomial’)
```

1C. The following is the full summary of the fitted logistic regression model from above.

```
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)    
## (Intercept)  -7.570e+01  1.132e+01  -6.687 2.28e-11 ***
## income       -1.013e-05  1.764e-05  -0.575    0.566    
## credit_score  1.171e-01  1.746e-02   6.707 1.98e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 301.05  on 349  degrees of freedom
## Residual deviance:  93.45  on 347  degrees of freedom
## AIC: 99.45
```

Are income and credit score statistically significant predictors for the approval status of the loan application? Explain in 1-2 sentences why or why not, and circle above any quantities relevant to your answer. (4 points)

Based on the summary table, income is not a statistically significant predictor as the corresponding p-value is 0.566, which is much greater than the 5% significance level. However, credit score is indeed a statistically significant predictor because the p-value is $1.98 * 10^{11}$, which is much lower than the 5% significance level.

1D. To assess the classifier's performance, we will validate the model on the testing set. To do this, please complete the blanks below (2 points).

```
# Predict probabilities on the testing set
predicted_probs <- predict(logistic_model,
                           newdata = loan_test,
                           type = "response")
```

1E. Consider the code below. Please explain in 2-3 sentences what this code accomplishes and why we would need it in order to use a logistic regression model to accomplish a classification task. (4 points)

```
predicted_labels <- ifelse(predicted_probs > 0.5, 1, 0)
```

The code converts predicted probabilities into binary class labels (1 or 0). In logistic regression, the output is a probability that ranges between 0 and 1, and we need to convert these probabilities into a binary class to make a final decision (e.g., approve or deny a loan). This code does so by setting a threshold at 0.5; probabilities above this threshold are classified as 1 (loan approval), and those below are classified as 0 (loan denial).

1F. Now we are going to change the model a little bit. Suppose that the parameter coefficient estimate for the intercept is -7.57, the parameter coefficient estimate for income is -0.013. And suppose that the parameter coefficient estimate for credit score is 0.772. Using these three coefficient estimates, calculate the probability of a loan application being approved when the applicant's annual income is $35000 with a credit score of 600. (10 points).

$$P(\text{ approval }) = \frac{1}{1 + e^{-(7.57 - 0.013 \times \text{ } 35000 \text{ } + 0.772 \times \text{ } 600 \text{ })}} = 0.652$$

1G. Suppose that you have also trained a Support Vector Machine (SVM) classifier for the loan dataset and you want to compare the performance of the two models. To assess their performance, you calculate the confusion matrix for both models and obtain the following.

Classifier	Accuracy	Sensitivity	Specificity	F1 score
Logistic	0.65	0.51	0.83	0.59
SVM	0.73	0.53	0.85	0.64

Based on the table above, which classifier appears to be performing better, and why do you think it is performing better? (3 points)

The SVM classifier appears to perform better. Based on the performance metrics table, the SVM classifier has a higher accuracy, which indicates a higher overall rate of correct predictions. Additionally, it has better sensitivity (or recall) and specificity, meaning it more accurately identifies positive and negative cases. The F-1 score, which is a balance between precision and recall, is also higher for SVM, suggesting it's more balanced in terms of false positives and false negatives compared to the logistic regression model.

1H. Consider a different problem now. Suppose there is a horse race, and only two horses are racing, Sea Biscuit and Secrtariat. You are told that the odds of Sea Biscuit winning the race are 1:5. Using this information, please calculate the probability of Secretariat winning the face? (5 points).

The probability of Secretariat winning the race, given that the odds of Sea Biscuit winning are 1:5, is approximately 83.33% or $\frac{5}{6}$.

Problem 2: Model Analysis and Assessment (35 Points)

Suppose that we have trained a certain classifier and obtained the following confusion matrix. Calculate the following quantities. Please show your calculations, and round your answer to 3 decimal places. Note: '1' represents the positive class.

	Predicted 0	**Predicted 1**
Actual 0	47	19
Actual 1	8	26

2A. What is the Type I (False Positive) error rate? (3 points)

Type I error rate $= \frac{19}{47+19} = 0.288$

2B. What is the Type II (False Negative) error rate? (3 points)

Type II error rate $= \frac{8}{8+26} = 0.235$

2C. Calculate the sensitivity. (3 points)

sensitivity $= \frac{26}{26+8} = 0.765$

2D. Calculate the specificity. (3 points)

specificity $= \frac{47}{47+19} = 0.712$

2E. Calculate the precision. (3 points)

precision $= \frac{26}{26+19} = 0.578$

2F. Explain in 2-3 sentences what an ROC curve represents, how it relates to the confusion matrix, and why we use it to evaluate a model's quality. (6 points)

The ROC curve represents the performance of a classification model at all classification thresholds. It shows the trade-off between sensitivity(the True Positive Rate) and specificity(False Positive Rate) at different classification thresholds and is used to evaluate the quality of a model because it provides a robust metric to assess the model's performance across all thresholds, highlighting how well the model can distinguish between classes.

2G. Explain in 1-2 sentences the advantage of using a random forest classification model instead of a single decision tree. (4 points).

Random forests are an ensemble method that combines the predictions of multiple decision trees, which tends to improve the overall predictive accuracy and control overfitting. Each tree is trained on a random subset of the data, making the ensemble less likely to be overfitted to the training data compared to a single decision tree.

2H. Explain in 1-2 sentences the advantage of using a support vector machine (SVM) model instead of a logistic regression model. (4 points).

The advantage of using a Support Vector Machine (SVM) model over a logistic regression model is that SVMs can efficiently perform a non-linear classification using the ”kernel” method, implicitly mapping their inputs into high-dimensional feature spaces. This makes SVMs particularly well-suited for classification problems with complex boundaries.

2I. Explain in 2-3 sentences the difference between a soft margin SVM and a hard margin SVM. (6 points).

The difference between a soft margin SVM and a hard margin SVM lies in how they handle outliers and noise. A hard margin SVM strictly maximizes the margin while allowing no misclassifications (perfect separation), which can lead to overfitting if the data has noise. A soft margin SVM allows some misclassifications while still maximizing the margin, introducing a slack variable to handle outliers and noise, thus providing a more generalized solution.

Problem 3: Neural Networks and Reinforcement Learning. (15 points)

3A. What is forward propagation in the context of neural networks? (2 points)

a). The process of adjusting weights based on error rates

b). The flow of information from the input to the output layer

c). The method of selecting the best activation function

d). The technique of dividing data into training and test sets

3B. In RL, what of the following best describes the sequence followed by an agent? (2 points)

a). Observes a state, takes an action, leads to state transition, receives a reward.

b). Receives a reward, takes an action, leads to state transition, observes a state.

c). Takes an action, observes a state, leads to state transition, receives a reward.

d). Observes a state, receives a reward, leads to state transition, takes an action.

3C. True or false: A neural network is a semi supervised algorithm. (2 points)

False

3D. True or False: A neural network with one hidden layer can approximate many non-linear functions. (2 points)

True

3E. True or False: A random process in which the probability of each state is independent of every other state is an example of a Markov Process. (2 points)

False

3F. Please explain in 2-4 sentences what is the Universal Approximation Theorem and why it is important in machine learning. (5 points)

The Universal Approximation Theorem states that any continuous function can be realized by a network with one hidden layer (given enough hidden neurons) on compact subsets of R^n. This theorem is important in machine learning because it provides a theoretical guarantee that neural networks can model any complex function, given enough neurons and proper training. This forms the foundational basis for using neural networks in a wide range of applications.

Problem 4: Large Language Models. (15 Points)

Suppose we have a dictionary containing words related to a movie: [‘Tom’, ‘said’, ‘the’, ‘movie’, ‘exciting’, ‘Mary’, ‘is’, ‘liked’, ‘and’]. We want to train a Language Model (LLM) and we need to start with a vector embedding of some phrases.

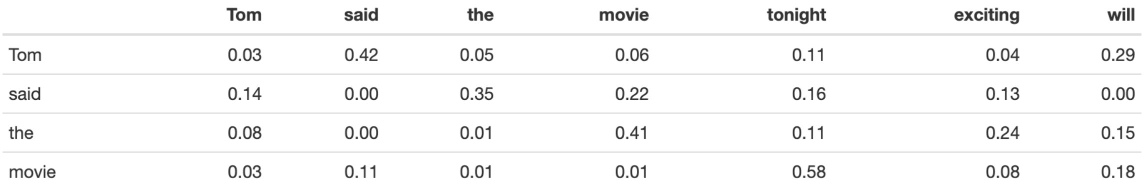
4A. Using one-hot embedding, as we learned in class, please represent the phrase “Tom said the movie is exciting” in its vector notation. (4 points)

Tom: [1, 0, 0, 0, 0, 0, 0, 0, 0]
said: [0, 1, 0, 0, 0, 0, 0, 0, 0]
the: [0, 0, 1, 0, 0, 0, 0, 0, 0]
movie: [0, 0, 0, 1, 0, 0, 0, 0, 0]
is: [0, 0, 0, 0, 0, 0, 1, 0, 0]
exciting: [0, 0, 0, 0, 1, 0, 0, 0, 0]

4B. Briefly explain in 1-2 sentences the purpose of a Multi-Layer Perceptron (MLP) in the context of LLMs. (3 points)

The purpose of a Multi-Layer Perceptron (MLP) is to introduce non-linearity into the process, which is crucial for capturing the complex patterns in language that linear operations cannot.

4C. Suppose we have reached the Unembedding step in the following (slightly shorter) dictionary. The final prediction matrix we generated is given as follows.



	Tom	said	the	movie	tonight	exciting	will
Tom	0.03	0.42	0.05	0.06	0.11	0.04	0.29
said	0.14	0.00	0.35	0.22	0.16	0.13	0.00
the	0.08	0.00	0.01	0.41	0.11	0.24	0.15
movie	0.03	0.11	0.01	0.01	0.58	0.08	0.18

Based on the table, what do you think the model would predict as the next token given the first two words “Tom said . . . ”? (4 points)

the

4D. Now, given the sentence “Tom said the movie . . . ”, and using again the table from problem 4C, what will be the next token predicted, and what is the probability associated with this prediction? (4 points)

tonight; 0.58